下载APP

08 | 案例串讲：从0开始，搭建一个规范的A/B测试框架

2020-12-24 张博伟

A/B测试从0到1

进入课程 >

讲述：张博伟

时长 15:42　大小 14.39M

你好，我是博伟。

经过前面几节课的学习，相信你不仅掌握了做 A/B 测试的统计原理，还知道了一个规范的 A/B 测试的流程是什么样的，以及关键步骤中有哪些需要注意的地方。

今天这节课的内容，整体来说不会太难，主要是用一个音乐 App 提升留存率的案例，来串讲一下我们学过的统计知识，以及做 A/B 测试的几个核心步骤。

在学习这节课的过程中，一方面，如果你还有一些没有完全搞懂的内容，可以再针对性复习下，查漏补缺；另一方面，之前几节课的内容容量都比较大，今天的案例串讲相当于帮助你理清思路，清空大脑，然后再有效地去吸收进阶篇的知识。

好了，那我就通过下面音乐 App 这个案例，来带你走一遍流程。

从业务问题出发，确定 A/B 测试的目标和假设

咱们今天案例里的产品是一款音乐 App，用户只要每月付费就可以免广告畅听千万首音乐。当然，除了最基本的播放音乐功能，产品经理还给这款 App 设计了很多便利的功能，比如用户可以把喜欢的音乐加入收藏夹，可以创建不同的歌单，还可以离线下载以便随时随地畅听自己喜欢的音乐，等等。

数据科学家通过数据分析也发现，使用这些便利功能的用户往往有着高于平均水平的续订率，说明这些便利功能确实有助于提升用户留存。但是也有一个问题一直困扰着团队：这些功能虽然方便实用，有助于优化用户的听歌体验，但是使用率却一直不高。使用率不高，从长期来看，势必会影响用户留存。

团队通过用户调研才发现其中的原因。

由于 App 的页面设计崇尚简洁，这些功能一般就存放在每首歌曲的功能列表中，而用户往往需要点击两次才能使用：第一次先点击功能列表，第二次再点击具体的产品功能。一方面，很多用户，尤其是新用户，并没有发现这些功能。另一方面，点击两次才能使用，用户体验并不好，慢慢地也就退订了。

那么，我们现在的目标就非常明确了：**增加用户对产品功能的使用率。**

如何增加这个使用率呢？你可能会说，把每个功能都直接显示出来，让用户一目了然，不就可以提高它们的使用率了嘛！产品经理刚开始就想到了这一点，但是后来发现功能太多，全部直接显示出来，会让歌曲界面看起来非常杂乱，会让用户体验更糟糕。

既然产品交互界面的改动被否定了，那么我们可不可以主动告知用户这些功能怎么使用呢？

比如说，在新用户刚注册登录后就告知他们每个功能的用法。不过这个想法很快也被产品经理否定了，毕竟新用户刚登录时并不会用到所有功能。这很好理解，因为没有需求嘛，新用户在看到这些功能时肯定也没有什么反应，所以新用户在第一次登录时一般都会跳过产品功能介绍。

之前的 A/B 测试也验证了这一点。只有在用户有使用这个功能的需求时，再告知他们，才最有效果。

于是团队的假设就是：**在用户有需求时，通过弹窗的形式告知用户相关使用功能，以此提升相关功能的使用率**。这样的话，既能避免对每一个新用户的打扰，又能满足有需求的用户，两全其美。

确定 A/B 测试的评价指标

确定了目标和假设之后，就可以开始定义评价指标了。

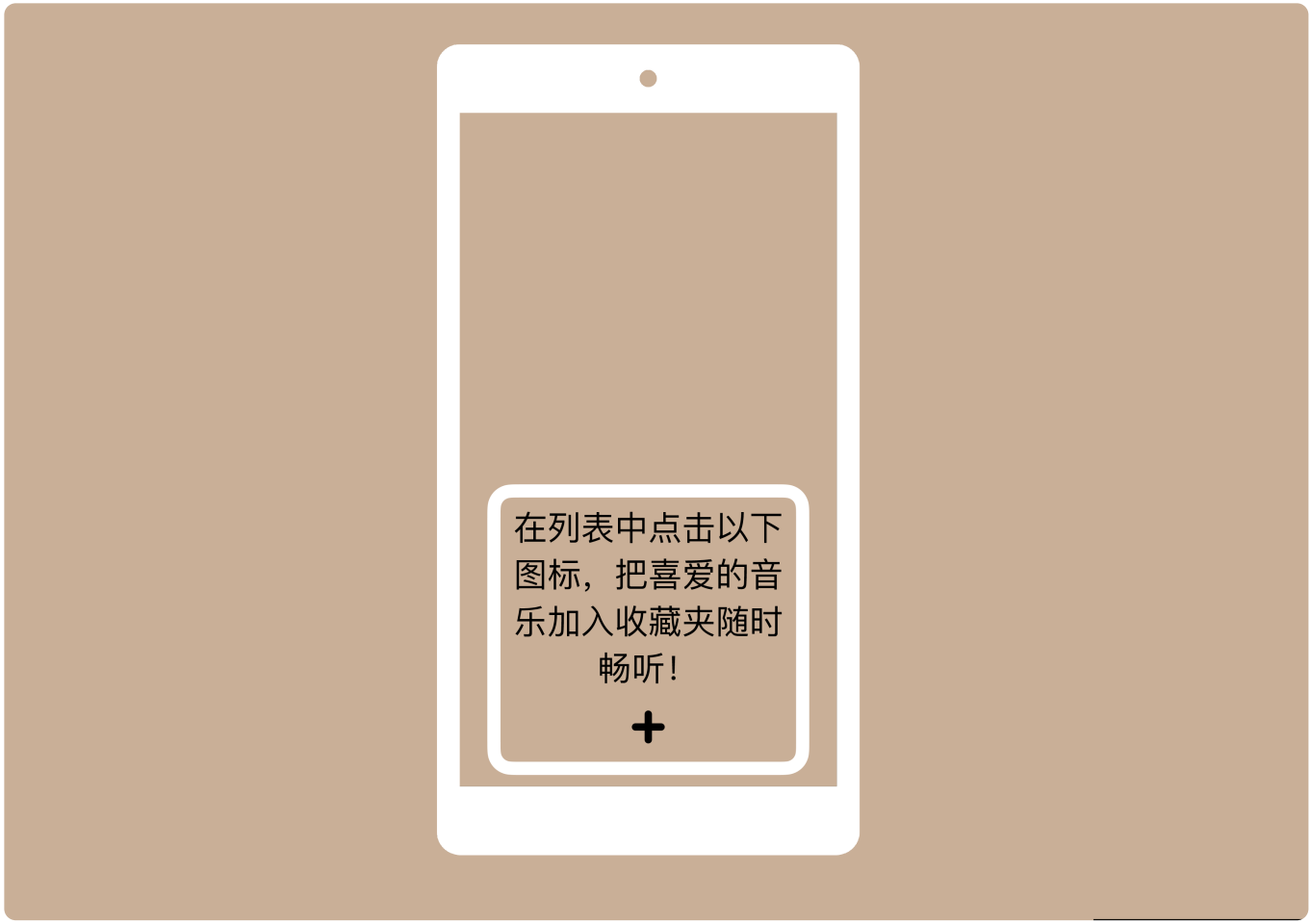
团队准备先拿“把喜欢的音乐加入收藏夹”这个功能来做一个 A/B 测试，验证以上的假设是否成立。

因为要在用户有需求的时候再告知用户，所以我们就需要一个条件来触发这个告知。那么，我们的首要任务就是**确定触发条件**：只有当用户从来没有用过这个功能（如果用户知道这个功能的话我们就没有必要告知了），并且播放同一首歌曲达到 x 次时（以此来判断用户对某首歌曲的喜爱程度），我们才会给用户发送弹窗通知。

经过数据科学家的数据分析，最终确定了 x 的最优值为 4，所以该功能的弹窗的最终触发条件为：

> 该用户从来没用过“把喜欢的音乐加入收藏夹”这个功能。
>
> 该用户已经对某首歌听了 4 次，当播放第 5 次时触发弹窗。

需要说明的是，因为弹窗是为了要告知用户，不需要重复提醒，所以每个符合触发条件的用户也只能收到一次，不能多次触发。



在这个 A/B 测试中把用户随机分为实验组和对照组，每组 50%。

在实验组中，如果用户满足了触发条件，系统就会发送弹窗提醒（如上图）。

在对照组中，用户不会收到弹窗提醒，不管是否符合触发条件。

确定了目标和假设，现在我们来具体定义下评价指标：

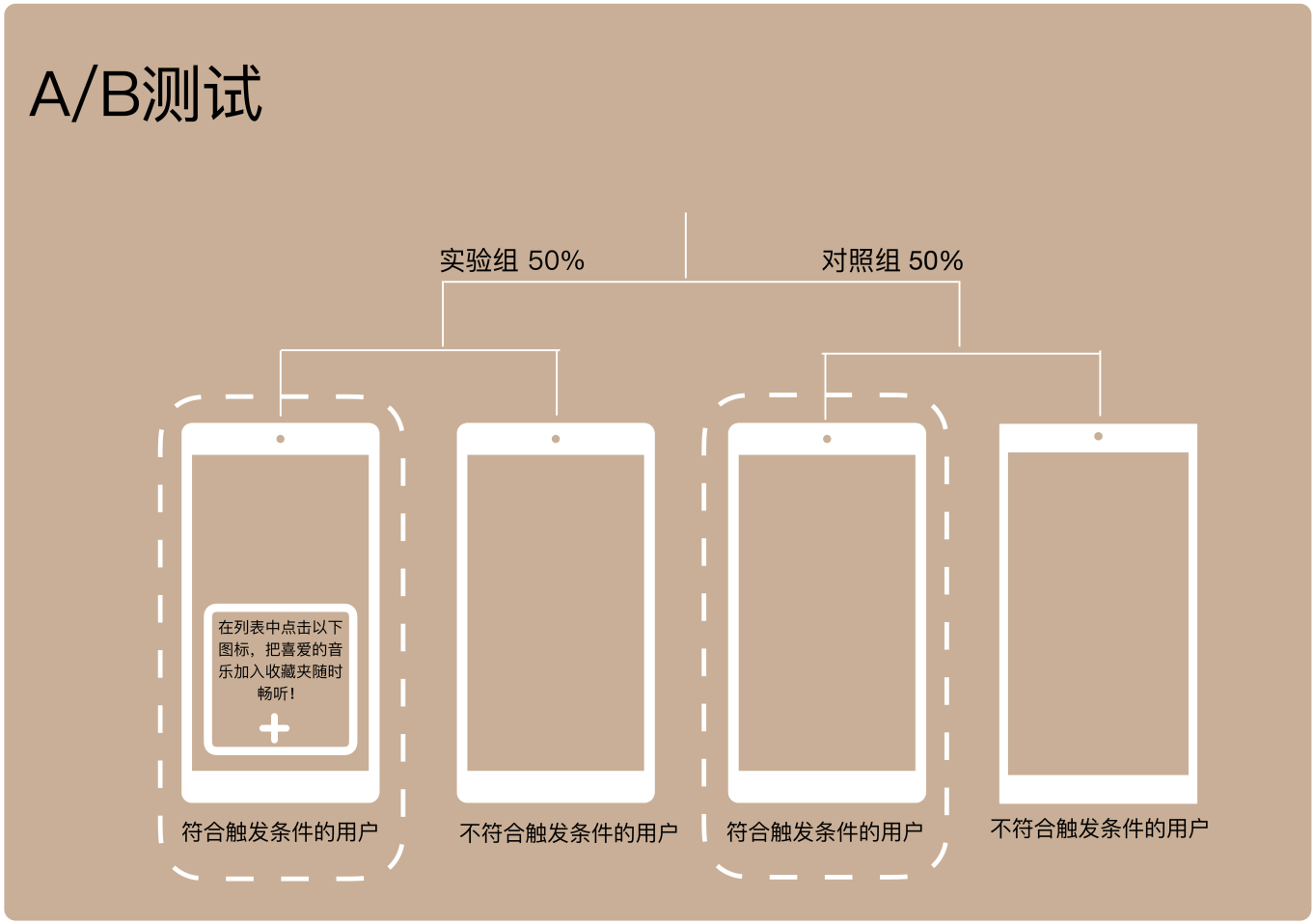
“把喜欢的音乐加入收藏夹”功能的使用率 = 使用了“把喜欢的音乐加入收藏夹”的用户总数 / 实验中的用户总数。

很明显，这是一个概率类的指标，也就是说在实验中的这些用户，使用了“把喜欢的音乐加入收藏夹”这个功能的概率有多少。不过，为了使我们的评价指标更加具体，也方便之后的计算，所以这里我们要搞清楚两个问题。

第一个问题，如何定义“实验中的用户”？

鉴于用户只有满足了条件才会触发弹窗，并不是所有在实验中的人都会受到影响，所以测试时不能用所有被分配在实验中的用户，因为这样就会引入没有受到影响的用户（那些被分配在实验中但是却没有满足触发条件的用户），从而降低测试的准确性。所以一定要注意，这里的“实验中的用户”应该是符合触发条件的用户（下图中虚线部分）。

在实验组中就是触发弹窗的用户，在对照组中则为符合触发条件的用户（因为对照组中的用户不管符合不符合触发条件都不会触发弹窗）。



第二个问题，如何确定用户从触发弹窗开始到最终使用功能的时间窗口期呢？

因为本次 A/B 测试是要检测弹窗是否会对相关功能的使用率有所提升，而且每个用户触发弹窗的时间不同，所以需要事先规定一个统一的时间窗口期来衡量，比如触发后 x 天之内的使用率，这样统一化是为了使指标更加清晰准确。

因为弹窗告知在这里具有及时性，及时性也就是说在用户有需求时，所以如果用户是受到弹窗的影响才使用相关功能时，肯定会在看到弹窗不久后就使用了。我们这里就把 x 设为 1，即触发后 1 天内的使用率。

搞清楚了这两个问题，我们就可以把评价指标最终定义为：

“把喜欢的音乐加入收藏夹”功能的使用率 = 在符合触发条件后 1 天之内使用了“把喜欢的音乐加入收藏夹”的用户总数 / 实验中的符合触发条件的用户总数

光确定评价指标的具体定义还不够，为了更了解咱们的评价指标，得出准确的实验结果，我们还要从统计的角度来看下这个指标的波动性如何。

通过对历史数据的回溯性分析，得到了用户在符合触发条件后一天之内使用相关功能的平均概率为 2.0%，通过统计公式最后求得该指标 95% 的置信区间为[1.82%，2.18%]。这就说明如果测试结束后两组评价指标的值均落入这个波动范围内，则说明两组无显著不同，属于正常波动范围。

选取实验对象的单位

确定了 A/B 测试的评价指标后，接下来我们要确定下实验对象的单位了。

因为本次实验的弹窗是用户可见的变化，而且评价指标是以用户为单位，**所以我们选择用户作为最小实验对象单位，具体来说，可以选用用户 ID**，因为这些用户必须登录后才能享受音乐服务。

计算所需的样本大小和实验所需时间

我们继续往下走，就该计算实验所需的样本量了。这里，我们需要先确定 4 个统计量：

显著水平（Significance Level）α。

Power （1 – β）。

实验组和对照组的综合方差 σ^2_{pooled}。

实验组和对照组评价指标的差值δ。

一般 A/B 测试中显著水平默认为 5%，Power 默认为 80%， 我们的案例也遵循这样的原则。至于两组评价指标之间的差值，根据我们之前算出的波动性，两者的差值要在 0.18% 以上，才算是统计显著的变化，那么我们就取 0.2%。至于综合方差，因为是概率类的指标，我们就可以用统计公式直接算出。

确定了这些统计量后，我们算出实验组和对照组各需要至少 8.07 万个符合触发条件的用户，一共需要 16.14 万用户。而数据分析显示每天符合触发条件的新用户大约为 1.7 万人，所以本次实验大约需要 10 天时间完成。

那么当我们完成了对整个 A/B 测试的设计工作后，现在就让测试跑起来，收集数据，等到样本量达到预期时就开始分析测试结果。

分析测试结果

经过了一周多的等待，我们的样本量终于达标，可以来分析最终的结果啦。不过在分析结果前，我们还要确保 A/B 测试在具体实施过程中符合我们最初的设计，保证测试的质量品质，这时候就要做合理性检验。

我们用最常见的护栏指标来做检验。

实验 / 对照组样本大小的比例是否为 50%/50%。

实验 / 对照组中特征的分布是否相似。

经过分析发现，本次 A/B 测试完全通过了这两项护栏指标的合理性检验，说明试验实施的正如预期。

那么，现在我们就开始正式分析实验结果了。

实验组：样本量为 80723，符合触发条件一天之内使用功能的用户为 3124，使用率为 3.87%。

对照组：样本量为 80689，符合触发条件一天之内使用功能的用户为 1598，使用率为 1.98%。

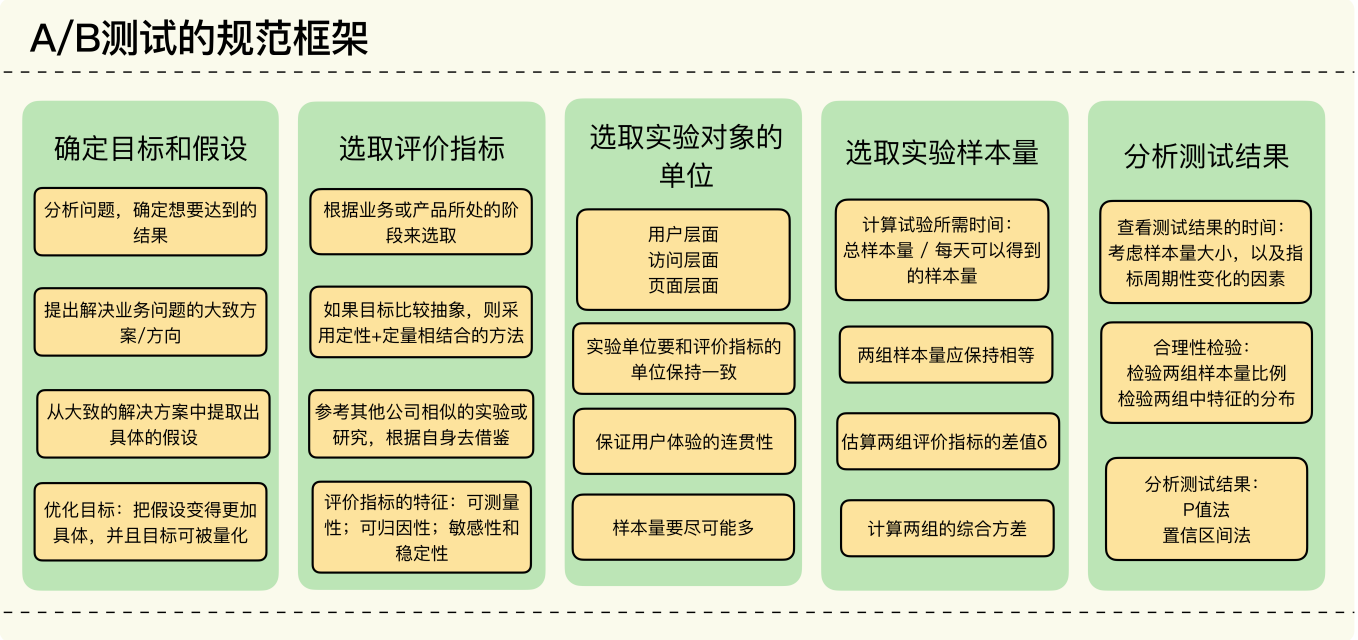
```
        2-sample test for equality of proportions with continuity correction

data:  c(3124, 1598) out of c(80723, 80689)
X-squared = 506.69, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.01724190 0.02054972
sample estimates:
    prop 1     prop 2
0.03870025 0.01980443
```

根据结果我们得到的 P 值接近于 0 而且远远小于 5%，同时我们计算出两组评价指标差值的 95% 的置信区间为[1.72%，2.05%]，不包括 0，说明这两组的使用率显著不同，事实上实验组的使用率几乎等于对照组的两倍，证明了在用户需要时的弹窗提醒确实有效果！

得到这个振奋人心的结果后，团队决定把“把喜欢的音乐加入收藏夹”功能的弹窗提醒推广到所有符合触发条件的用户，同时也计划对其他功能的弹窗做类似的 A/B 测试，来验证它们的效果。如果一切进行顺利的话，就将这些弹窗全部推广，长期来看肯定会增加用户的留存率！

小结

通过这个案例串讲，你肯定对做 A/B 测试的关键步骤有了更具体、更深层次的认识了。



那么基础篇的内容到这里也就结束了。接下来我们就会进入到进阶篇的学习。

在进阶篇，我会给你讲解更多偏经验和方法论的知识。针对做 A/B 测试时经常出现的一些问题，我会给你讲解它们的成因，给出解决办法。针对面试中常出现的一些考点，我会结合我做面试官的经验，来给你一些解题思路。

最后我还想强调一下，学习这件事本来就是反复和持续的。进阶篇的内容会和基础篇有不少联系。所以在学习进阶篇的课程时，我也希望你能够不断温习、思考之前学过的知识。待课程结束，再回头看基础篇这些内容，相信你会有一种“蓦然回首，原来 A/B 测试如此简单”的畅快感和收获感。

思考题

回忆你之前做过或者经历过的 A/B 测试，它们是否有这些基本的流程步骤？如果缺少的话，是缺少哪些步骤，为什么？如果还有其他步骤，也和我分享一下吧。

如果你学完今天的案例串讲，对 A/B 测试的流程、步骤有了更清晰的认识，欢迎你点击“请朋友读”，把今天的内容分享给你的同事、好友，大家一起学习、成长。好，感谢你的收听，我们进阶篇的课程再见。

提建议

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 07 | 分析测试结果：你得到的测试结果真的靠谱吗？

下一篇 09 | 测试结果不显著，要怎么改善？

精选留言 (10)

写留言

贤者时间
2020-12-24

请问老师，有这么一个场景，电商平台对低活的用户发放优惠券，优惠券需要用户手动领取，想做ab测试知道发放优惠券是否有提升活跃的效果，那么这里的实验组应该是所有发放了优惠券的用户，还是领取了优惠券的用户呢？

展开

作者回复: 你好，这里我的理解是并不是所有发放了优惠卷的用户都会领取，用户需要一定的触发条件（比如要登录）才可以领取？这样的话计算指标时要用满足了触发条件的用户，而不是所有被发放的用户。
我会在第9节课中详细讲触发这种情况的，敬请期待。

2

孙小军
2020-12-31

请问老师，前面课程提到的A/A测试在这个案例里面应该怎么体现和实施呢？

作者回复: 你好，A/A测试的话一般是在做A/B测试前进行，具体实施就是对流量随机均分，不过两组的体验完全相同，然后去看评价指标的波动性（比如两组指标是否相同，如果不同是否显著等），从而了解其稳定性，如果两组有显著的不同，那就说明指标稳定性差，不适合做评价指标。

1

贤者时间
2020-12-28

张老师，请问做AB测试，其中从指标选取到分析结果的这些流程，都是在公司搭建的AB测试平台自动化完成的吗？我感觉其中很多步骤都是和业务强相关的，很难标准流程化。比如每个ab测试选的指标不一样的话，那么指标的计算也要重新定义。那么AB测试平台在全流程中，承担的是哪些部分呢？

展开

作者回复: 你好，整个这个过程不全是，就像你所说的，流程中的很多步骤都是和业务息息相关的，很难完全标准化，不过可以对常用指标做到部分标准化，比如针对拉新方面的各种A/B测试，那么新用户转化率肯定是一个常用指标，再比如对于优化用户参与度方面的各类A/B测试，使用时间/频率肯定是常用指标等等。A/B测试平台在全流程中，主要承担具体实施A/B测试这个功能（包括如何随机分流，如果控制前端给到不同的用户体验等等），有的会包括部分标准化指标的计算，还有最后的结果计算等。

2

KayeArt
2020-12-25

请问老师，在做决策时，是否总是要考虑实际显著性呢？在什么情况需要考虑，什么情况下不需要考虑？
如果不考虑实际显著性，是不是意味着只要试验结果在统计上是显著的，就证明实验组和控制组是有显著差异的？

展开

作者回复: 你好，1. 如果是改变现有的状态且改变有一定成本时需要是要考虑实际显著性的，但是对于探索性质的且改变成本不高的情况下可以不考虑，比如你想看看改变不同广告内容看广告效

果如何，这个时候改变不同的广告内容可能只是在A/B测试平台组中进行简单设置，其实并没有很高的成本。2. 如果不考虑实际显著性的话就以统计显著为准。

1 1

那时刻
2020-12-24

请问老师，该用户已经对某首歌听了 4 次，当播放第 5 次时触发弹窗。这个4次是怎么统计得来的呢？是依据历史用户点击收藏夹平均听歌次数么？
另外，10天的试验周期，是否考虑用户数据随日期波动性呢？比如是否跨周末，节假日的影响。

展开

作者回复: 你好，1.对的！ 2. 如果指标有周期性波动是要考虑的，常见的周期性是一周，所以10天的话已经包含了一个周期了。

1

Han
2021-01-04

老师： “把喜欢的音乐加入收藏夹”功能的使用率 = 在符合触发条件后 1 天之内使用了“把喜欢的音乐加入收藏夹”的用户总数 / 实验中的符合触发条件的用户总数， 这里实验的用户数是80723，这个实验中的数据符合正态分布么？
我感觉不符，中心极限定理说样本量不低于30，每个样本量中包含N个样本。实验组用户数是80723，算是一个样本量吧？ 抽样的样本量应该至少30，是怎么满足的？

展开

作者回复: 你好，这里是概率类指标，所以判断标准有所改变（参看第一节课）：min(np,n(1-p)) >= 5。其中，n 为样本大小，p 为概率的平均值。

Geek_d4ccac
2020-12-30

请问老师, 文中”实验组：样本量为 80723，符合触发条件一天之内使用功能的用户为 3124，使用率为 3.87%” 这里的样本量80723是实验组中符合触发条件的用户数么? 对照组的样本量也是符合触发条件的用户数？ 感谢!

展开

作者回复: 嗯嗯你对两个问题的理解都是对的！

aiorosxu
2020-12-26

感觉对用户听过几次歌曲才触发弹窗这里，也可以进行A/B测试

作者回复: 对的！实践中确实是可以的！本文中的案例是先通过数据分析出来一个阈值，然后根据这个来触发弹窗，是为了让这个功能快速迭代出来，证明有效之后再进行像你说的这种微调各种参数的A/B测试。

西西
2020-12-25

想问个A/B测试之前的产品思路问题，案例中说到便利功能的使用与续订率相关，但便利功能的使用度可能与收听歌曲数正相关，但并不知道哪个与续订率存在因果关系。我们是不是也可以做实验去促进用户多收听歌曲呢？究竟针对哪项功能去做改进才是合理的呢？

展开

作者回复: 你好，首先要搞清楚两个变量之间的关系可以通过A/B测试或者其他因果推断的方法（比如PSM，我在第12节课中会讲解）。不过你说的这个情况，从实际的数据分析来看收听时间/频率这些指标一般都是会和续订率呈正相关的，不过优化这些指标是一个系统工程，包括了优化内容/推荐/产品体验/推送等等，是一个系统工程，本文中的例子是从产品功能这个具体领域入手，进行优化，我们当然可以在其他领域做A/B测试去优化嘛。其实更完整的因果链应该是：不同领域的优化->收听时间/频率增加->用户留存增加。

DataCenter 1
2020-12-25

老师您好，我对这里的实验&对照样本范围感到有疑问，在文中，比如实验组，样本范围是包括
1、符合触发条件的用户
2、不符合触发条件的用户
这两部分的用户构成了实验组用户群；...

展开

作者回复: 你好，你的理解和文中讲的是一样的呀。
文中（“把喜欢的音乐加入收藏夹”功能的使用率 = 在符合触发条件后 1 天之内使用了“把喜欢的音乐加入收藏夹”的用户总数 / 实验中的符合触发条件的用户总数）里的”实验中“就是包括对照组和实验组的。